Oracle Inequalities for Support Vector Machines that are based on Random Entropy Numbers

Ingo Steinwart
Information Sciences Group, CCS-3
Mail Stop B256
Los Alamos National Laboratory
Los Alamos, NM 87545
Tel: (505) 665-7914
Fax: (505) 667-1126
ingo@lanl.gov

June 11, 2009

Abstract

In this paper, we present a new technique for bounding local Rademacher averages of function classes induced by a loss function and a reproducing kernel Hilbert space (RKHS). At the heart of this technique lies the observation that certain expectations of random entropy numbers can be bounded by the eigenvalues of the integral operator associated to the RKHS. We then work out the details of the new technique by establishing two new oracle inequalities for support vector machines, which complement and generalize previous results.

AMS subject classification
primary: 68Q32,
secondary: 15A18, 41A46, 47B06, 68T05

Keywords
Statistical Learning Theory, Kernel-based Methods, Eigenvalues, Entropy Numbers

1 Introduction

Recent results [2, 14, 16, 19] establishing learning rates for support vector machines (SVMs) use Talagrand's inequality together with local Rademacher averages, see [1], to bound the estimation error, i.e., the statistical error of these learning methods. This approach requires to bound the local Rademacher averages of relatively complicated function classes that depend on both the loss function and the reproducing kernel Hilbert space (RKHS) used in the SVM. For this task, two approaches currently exists: The first one, which goes back to Mendelson [6] and is applied in [16, 19], uses Dudley's chaining together with uniform covering numbers of the RKHS, while the second one, applied in [2], uses another result by Mendelson [8] to

bound the Rademacher averages by the eigenvalues of the integral operator associated to the kernel of the RKHS. Currently, both approaches have advantages and disadvantages. For example, compared to uniform covering numbers, the eigenvalues are closer related to the learning problem at hand and provide, in general, a weaker notion of the complexity of the RKHS. In particular, the compactness of the input space is, in general, superfluous when using eigenvalues instead of uniform covering numbers. On the other hand, the analysis based on the eigenvalues is substantially more involved, and so far it is unclear whether, apart from a relatively simple case considered in [2], it can be carried out for more general settings. In addition, it remains so far unclear whether the analysis based on eigenvalues produces artifacts, such as the need of a quite restrictive noise assumption on the data-generating distribution. Consequently, it seems fair to say that currently neither of these two approaches are silver bullets.

In this paper, we present a new technique for bounding the local Rademacher averages, which combines the advantages of both approaches and simultaneously lacks their disadvantages. At the heart of our approach lies the simple observation that in Dudley's chaining argument one can use the functional inverse of covering numbers, i.e., entropy numbers. As a result, see Theorem 3.5, one can then bound the local Rademacher averages by the expectation of random entropy numbers. In the past, see e.g. [17], these in turn have been bounded by uniform entropy (or covering) numbers, which led to the first approach discussed above. To overcome the disadvantages of this approach, we use a result that bounds these random entropy numbers by the eigenvalues of the associated integral operator. In a nutshell, our new technique thus uses certain properties of entropy numbers to go from complicated functions classes considered in local Rademacher averages to scaled balls of RKHSs, and then uses specific features of RKHSs to make the step from random entropy numbers to eigenvalues.

We illustrate how to use this new technique by deriving two new oracle inequalities for SVM type methods, which both use eigenvalues estimates as a complexity measure for the RKHSs. To be more precise, the first oracle inequality considers classical SVMs, while the second one deals with an SVM type approach that uses a lighter regularization term. We further show that both results nicely complement and generalize corresponding findings from [14, 2]. In particular, it turns out that the new oracle inequalities combine the advantages of the two approaches discussed above while simultaneously lacking their disadvantages.

The rest of this paper is organized as follows. In Section 2, we first explain our new approach in more detail and provide some results that connect random entropy numbers to eigenvalues. We then present and discuss the two oracle inequalities mentioned above. The proofs of these inequalities can be found in Section 3. Finally, we have attached an appendix that contains the relatively elegant, yet unusual proof for the connection between random entropy numbers and eigenvalues.

2 Main Results

In the following, X always denotes a measurable space that is equipped with some probability measure μ. Moreover, H denotes a RKHS over X, whose kernel k :

$X \times X \to \mathbb{R}$ is assumed to be measurable. Let us further assume that it satisfies

$$\|k\|_{L_2(\mu)} := \left(\int_X k(x,x)\, d\mu(x) \right)^{1/2} < \infty \,.$$

Then it is well-known, see e.g. [13, Chapter 4.3.], that H consists of square integrable functions and the inclusion $\mathrm{id} : H \to L_2(\mu)$ is continuous with $\|\mathrm{id} : H \to L_2(\mu)\| \leq \|k\|_{L_2(\mu)}$. Moreover, the integral operator $T_{k,\mu} : L_2(\mu) \to L_2(\mu)$ defined by

$$T_{k,\mu} g(x) := \int_X k(x,x') g(x') d\mu(x') \,, \qquad g \in L_2(\mu),\, x \in X \,, \tag{1}$$

is known, see e.g. again [13, Chapter 4.3.], to be self-adjoint, positive, and compact. In addition, its ordered sequence (with geometric multiplicities) of eigenvalues $(\lambda_i(T_{k,\mu}))_{i \geq 1}$ is summable, i.e.,

$$\sum_{i=1}^{\infty} \lambda_i(T_{k,\mu}) < \infty \,.$$

As already mentioned in the introduction, it has been shown in [2] that the speed of convergence of $\lim_{i \to \infty} \lambda_i(T_{k,\mu}) = 0$ can be used to determine learning rates for SVMs using the hinge loss. In particular, [2] showed that faster rates of convergence result in faster learning rates. Of course, the behavior of the eigenvalues depends, in general, not only on the kernel k but also on the measure μ, which for learning problems equals the marginal distribution P_X of the data-generating distribution P on $X \times Y$, where $Y \subset \mathbb{R}$ is the set of possible labels. Therefore, the result in [2] seems to make it possible to identify distributions P_X for which SVMs learn particularly fast. Unfortunately, however, the results in [2] only hold under a restrictive form of the sharpest Tsybakov noise assumption, see below for the details, and hence they cannot be used to explain the learning behavior of SVMs in realistic settings.

Another, more classical way to determine learning rates for SVMs and other learning algorithms is based on the concept of covering numbers, or, as observed in [18], on entropy numbers, which are the "inverse" of covering numbers. Let us only recall the definition of entropy numbers since they have, as we will describe below, a tight connection to eigenvalues. To this end, let E and F be Banach spaces and $S : E \to F$ be a bounded linear operator. Then the (dyadic) entropy numbers $e_i(S)$, $i \geq 1$, of S are defined by

$$e_i(S) \;:=\; \inf\left\{ \varepsilon > 0 : \exists\, x_1, \ldots, x_{2^{i-1}} \in SB_E \text{ such that } SB_E \subset \bigcup_{j=1}^{2^{i-1}} \bigl(x_j + \varepsilon B_F\bigr) \right\},$$

where B_E and B_F denote the closed unit balls of E and F, respectively. Clearly, S is compact if and only if $\lim_{i \to \infty} e_i(S) = 0$, and the speed of this convergence can be considered as a measure on how compact S is. Now, if X is a compact space and k is continuous, then it is well-known that $\mathrm{id} : H \to C(X)$ is compact, and the convergence of the corresponding entropy numbers can be used to determine learning rates for SVMs, see [14, 16, 19]. Compared to [2], these learning rates hold for less restrictive assumptions on P, and are thus more widely applicable. On the

downside, however, the entropy numbers of $\mathrm{id} : H \to C(X)$ are *independent* of P_X, and therefore they do not give us the opportunity to identify marginal distributions P_X for which SVMs learn particularly fast.

As we will see in the proofs of our main results, it is, however, not necessary to use $C(X)$-entropy numbers in [14]. Instead, it will turn out that it suffices to use expectations of random entropy numbers. More precisely, if for given $D_X \in X^n$ we write D_X for the corresponding empirical measure, then the behavior of

$$\mathbb{E}_{D_X \sim P_X} e_i(\mathrm{id} : H \to L_2(\mathrm{D}_X)) \,, \qquad i \geq 1 \,, \tag{2}$$

can be used to determine oracle inequalities for SVMs, and thus learning rates. Unfortunately, however, expectations of random entropy numbers are known to be notoriously hard to deal with, which to some extend may explain why the expectation is often replaced by a supremum, see e.g. [17]. Obviously, the latter, presumably sub-optimal, approach could be avoided, if we could "move" the expectation inside the entropy numbers, that is, if we could consider $e_i(\mathrm{id} : H \to L_2(\mu))$, instead. Surprisingly, the following result shows that this is indeed possible:

Theorem 2.1 *Let k be a measurable kernel on X with separable RKHS H and μ be a probability measure on X such that $\|k\|_{L_2(\mu)} < \infty$. Assume that there exist constants $0 < p < 2$ and $a \geq 1$ such that*

$$e_i(\mathrm{id} : H \to L_2(\mu)) \leq a\, i^{-\frac{1}{p}} \,, \qquad i \geq 1 \,. \tag{3}$$

Then there exists a constant $c_p > 0$ only depending on p such that

$$\mathbb{E}_{D \sim \mu^n} e_i(\mathrm{id} : H \to L_2(\mathrm{D})) \leq c_p\, a\, i^{-\frac{1}{p}} \,, \qquad i, n \geq 1.$$

The proof of the theorem above yields constants c_p with $c_p \to \infty$ for $p \to 0$, but so far, it is unclear whether this is an artifact of our techniques. Moreover, the theorem clearly fails to provide a tight relationship, if $e_i(\mathrm{id} : H \to L_2(\mu))$ decreases with a rate faster than polynomial. For example, for a Gaussian RBF kernel with *fixed* width, it is known from e.g. [20] that the corresponding entropy numbers enjoy a certain exponential decay. In this case, Theorem 2.1 shows that the expected random entropy numbers decay with arbitrarily fast polynomial rate, but it fails to answer the question whether the expected random entropy numbers enjoy the same exponential decay. On the other hand, for SVMs based on Gaussian RBFs with *flexible* width, the sharpest existing statistical analysis in [16] uses bounds on the entropy numbers that only decrease polynomially but enjoy a better dependence on the used width of the kernel. Clearly, for such bounds, Theorem 2.1 produces the desired translation since the constant a, which in the Gaussian case depends on the kernel width, remains unchanged modulo the constant c_p.

Theorem 2.1 can be restated in terms of Lorentz sequence norms, see e.g. Chapter 1.5 in [3]. To do so, recall that for $p \in (0, \infty)$ and a decreasing, non-negative sequence (a_i) the Lorentz (p, ∞)-norm is defined by

$$\|(a_i)\|_{p,\infty} := \sup_{i \geq 1} i^{\frac{1}{p}} a_i \,.$$

Consequently, Theorem 2.1 states that, for all $0 < p < 2$, there exists a constant $c_p > 0$ such that

$$\big\|\big(\mathbb{E}_{D\sim\mu^n} e_i(\mathrm{id} : H \to L_2(\mathrm{D}))\big)\big\|_{p,\infty} \le c_p \big\|\big(e_i(\mathrm{id} : H \to L_2(\mu))\big)\big\|_{p,\infty} .$$

The following lemma shows that a similar relation holds between the eigenvalues and the $L_2(\mu)$-entropy numbers.

Lemma 2.2 *Let k be a measurable kernel on X with separable RKHS H and μ be a probability measure on X such that $\|k\|_{L_2(\mu)} < \infty$. Then for all $0 < p < 1$ there exists a constant $c_p > 0$ only depending on p such that*

$$c_p \big\|\big(e_i^2(\mathrm{id} : H \to L_2(\mu))\big)\big\|_{p,\infty} \le \big\|(\lambda_i(T_{k,\mu}))\big\|_{p,\infty} \le 4\big\|\big(e_i^2(\mathrm{id} : H \to L_2(\mu))\big)\big\|_{p,\infty} .$$

The lemma above basically states that the eigenvalues and the *squared* $L_2(\mu)$-entropy numbers have the same asymptotic behavior as long the eigenvalues do not decrease faster than polynomial. In particular, if we assume

$$\lambda_i(T_{k,P_X}) \le a^{\frac{1}{p}}\, i^{-\frac{1}{p}} , \qquad i \ge 1, \tag{4}$$

for some constants $a \ge 1$ and $0 < p < 1$, then Lemma 2.2 yields a constant $c_p > 0$ such that

$$e_i(\mathrm{id} : H \to L_2(\mu)) \le c_p\, a^{\frac{1}{2p}}\, i^{-\frac{1}{2p}} , \qquad i \ge 1,$$

and hence Theorem 2.1 shows

$$\mathbb{E}_{D\sim\mu^n} e_i(\mathrm{id} : H \to L_2(\mathrm{D})) \le \tilde{c}_p\, a^{\frac{1}{2p}}\, i^{-\frac{1}{2p}} , \qquad i \ge 1, \tag{5}$$

where $\tilde{c}_p$ is another constant only depending on p. As already indicated above, such an estimate can be used to bound the local Rademacher averages occurring in a statistical analysis based on Talagrand's inequality. Consequently, the implication from (4) to (5) provides a simple device to incorporate eigenvalue estimates into an analysis that enjoys the relative simplicity of the entropy number approach.

To illustrate this approach, we now present two resulting oracle inequalities for SVMs. To this end, we fix a nonempty compact set $Y \subset [-1, 1]$ and a probability measure P on $X \times Y$. Moreover, let H be a separable RKHS with *bounded* measurable kernel k satisfying

$$\|k\|_\infty \le 1 .$$

In addition, $L : Y \times \mathbb{R} \to [0, \infty)$ always denotes a continuous function that is convex in the second variable and satisfies $L(y, 0) \le 1$ for all $y \in Y$. Moreover, we assume that L is Lipschitz continuous in the sense of

$$\big|L(y, t_1) - L(y, t_2)\big| \le |t_1 - t_2| , \qquad y \in Y,\, t_1, t_2 \in \mathbb{R}. \tag{6}$$

In particular, we are interested in the hinge loss, which for $Y := \{-1, 1\}$ is defined by $L(y, t) := \max\{0, 1 - yt\}$, $y \in Y$, $t \in \mathbb{R}$. The function L will serve as loss function and consequently let us recall the associated L-risk

$$\mathcal{R}_{L,P}(f) := \mathbb{E}_{(x,y)\sim P} L(y, f(x)),$$

where $f : X \to \mathbb{R}$ is a measurable function. Note that our assumptions immediately give $\mathcal{R}_{L,P}(0) \leq 1$. Furthermore, the minimal L-risk is denoted by $\mathcal{R}^*_{L,P}$, i.e.

$$\mathcal{R}^*_{L,P} := \inf\{\mathcal{R}_{L,P}(f) \,|\, f : X \to \mathbb{R} \text{ measurable}\},$$

and a function attaining this infimum is denoted by $f^*_{L,P}$. In the following we always assume that there exists at least one such $f^*_{L,P}$. In addition, if there happens to be more than one such $f^*_{L,P}$, we assume that we have picked one *fixed* such function.

Recall that support vector machines, see [4, 10, 13], are based on the optimization

$$f_{P,\lambda} := \arg\min_{f \in H} \Big(\lambda \|f\|^2_H + \mathcal{R}_{L,P}(f) \Big), \tag{7}$$

where $\lambda > 0$ is a user-defined regularization parameter and the function $f_{P,\lambda}$ is known to be uniquely determined, see [13, Chapter 5.1]. Note that if we identify a training set $D = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ with its empirical measure, then $f_{D,\lambda}$ denotes the empirical estimator of the above learning scheme.

One way to describe the approximation error of SVMs is the 2-approximation error function

$$A_2(\lambda) := \lambda \|f_{P,\lambda}\|^2_H + \mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}^*_{L,P}\,, \qquad \lambda > 0,$$

which is discussed in some detail in [15] and Chapter 5.4 of [13]. In particular, the 2-approximation error function has a tight connection to the more classical approximation errors of the scaled unit balls $\lambda^{-1} B_H$. For a precise statement in this direction we refer to [13, Exercise 5.11].

With these preparations we can now formulate our first oracle inequality.

Theorem 2.3 *Let L, H, and P satisfy the assumptions above. Moreover, assume that there are constants $a \geq 1$ and $0 < p < 1$ such that*

$$\mathbb{E}_{D_X \sim P^n_X} e_i(\mathrm{id} : H \to L_2(\mathrm{D}_X)) \leq a^{\frac{1}{2p}}\, i^{-\frac{1}{2p}}\,, \qquad i \geq 1\,. \tag{8}$$

In addition, suppose that for all $0 < \lambda \leq 1$ and all $f \in \lambda^{-\frac{1}{2}} B_H$ we have

$$\mathbb{E}_P\big(L \circ f - L \circ f^*_{L,P}\big)^2 \leq c\,\big(\|f\|_\infty + 1\big)^{2-\vartheta}\,\big(\mathbb{E}_P(L \circ f - L \circ f^*_{L,P})\big)^{\vartheta} \tag{9}$$

for some constants $c \geq 1$ and $\vartheta \in [0,1]$. Then there exists a constant $K \geq 1$ only depending on c and p such that for all $0 < \lambda \leq 1$, $\varepsilon \in (0,1]$, $\tau \geq 1$, and $n \geq \tau$ satisfying $\varepsilon \geq A_2(\lambda) + \lambda$ and

$$\varepsilon \;\geq\; K\lambda^{-1} \max\left\{ \left(\frac{a}{n}\right)^{\frac{2}{2-\vartheta+\vartheta p}}, \left(\frac{a}{n}\right)^{\frac{2}{1+p}}, \left(\frac{\tau}{n}\right)^{\frac{2}{2-\vartheta}} \right\},$$

we have

$$P^n\Big(D \in (X \times Y)^n : \mathcal{R}_{L,P}(f_{D,\lambda}) - \mathcal{R}^*_{L,P} < A_2(\lambda) + \varepsilon\Big) \;\geq\; 1 - e^{-\tau}\,.$$

In principle it is possible to derive a value for the constant K from the proof of Theorem 2.3. However, we strongly believe that the proof does not provide a sharp value, and thus we omitted a detailed analysis.

To illustrate the theorem above let us now assume that L is the hinge loss. Moreover, assume that P is a distribution with Tsybakov noise exponent $q \in [0,\infty]$, i.e., there exists a $C > 0$ such that for $\eta(x) := P(y = 1|x)$, $x \in X$, and all $t > 0$ we have

$$P_X\big(\{x \in X : |2\eta(x) - 1| \le t\}\big) \;\le\; (C \cdot t)^q \,. \tag{10}$$

When $q > 0$, it follows from [16, Lemma 6.6] that the assumption (9) is satisfied with $\vartheta = \frac{q}{q+1}$ and $c = C^q + 2$. Moreover, it is simple to show the same is true when $q = 0$ but with $c = 5$. Let us further assume that the sample size n satisfies $n \ge a\tau$. Some easy estimates then show that the conditions on ε in Theorem 2.3 are satisfied if

$$\varepsilon \ge A_2(\lambda) + \lambda + K\lambda^{-1}\Big(\frac{a\tau}{n}\Big)^{\frac{2(q+1)}{q+pq+2}} \,, \tag{11}$$

that is, we have

$$\mathcal{R}_{L,P}(f_{D,\lambda}) - \mathcal{R}^*_{L,P} < 2A_2(\lambda) + \lambda + K\lambda^{-1}\Big(\frac{a\tau}{n}\Big)^{\frac{2(q+1)}{q+pq+2}} \tag{12}$$

with probability P^n not smaller than $1 - e^{-\tau}$. Now note that (8) is implied by the assumption

$$\sup_{D_X \in X^n} e_i(\mathrm{id} : H \to L_2(\mathrm{D}_X)) \le a^{\frac{1}{2p}}\, i^{-\frac{1}{2p}} \,, \qquad\qquad i \ge 1,$$

which was imposed in [14, Theorem 2.1 & Example 2.4]. Besides this, however, the oracle inequality in [14, Example 2.4] is identical to (12), and hence we see that in this sense Theorem 2.3 generalizes the results from [14]. Moreover, the implication (4) $\Rightarrow$ (5) shows that the oracle inequality of Theorem 2.3 also holds (modulo a constant depending only on p), if we replace the random entropy number assumption (8) by the eigenvalue assumption (4). Under the latter condition, [2] has also established an oracle inequality in the case that $x \mapsto \eta(x)$ is bounded away from 0, 1, and 1/2, that is, if a stronger version of (10) holds for $q = \infty$. However, their result becomes more interesting if the regularization term $\|\cdot\|_H^2$ in (7) is replaced by the lighter regularization $\|\cdot\|_H$. Interestingly, our techniques can also be used to derive an oracle inequality for such a regularization. To formulate the corresponding result, we define the 1-approximation error function

$$A_1(\lambda) := \inf_{f \in H}\Big(\lambda\|f\|_H + \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}\Big) \,, \qquad\qquad \lambda > 0, \tag{13}$$

which is based on this lighter regularization. Again, it is possible to show that there exists a unique minimizer $f^{(1)}_{P,\lambda}$ of the objective function in (13). In the following, we write $f^{(1)}_{D,\lambda}$ if P is an empirical measure based on the sample set D. In other words, $f^{(1)}_{D,\lambda}$ is the decision function produced by an algorithm using the lighter regularization. Moreover note that there is an intimate relationship between the new function A_1 and the 2-approximation error function. Indeed, [13, Exercise 5.11] can be used to show that, given a $\beta \in (0, 1]$, the following two conditions are equivalent:

i) There exists a constant $c > 0$ such that $A_2(\lambda) \leq c\lambda^{\beta}$ for all $\lambda > 0$.

ii) There exists a constant $\tilde{c} > 0$ such that $A_1(\lambda) \leq \tilde{c}\lambda^{\frac{2\beta}{1+\beta}}$ for all $\lambda > 0$.

In fact, the relationship between the constants c and $\tilde{c}$ can also be worked out modulo a universal constant, but for brevity's sake we omit the details. Let us now present our oracle inequality for this lighter type of regularization.

Theorem 2.4 *Let L, H, and P satisfy the assumptions above. Moreover, assume that both (8) and (9) are satisfied for some constants $a \geq 1$, $0 < p < 1$, $c \geq 1$, and $\vartheta \in [0, 1]$. Then there exists a constant $K \geq 1$ only depending on c and p such that for all $0 < \lambda \leq 1$, $\tau \geq 1$, and $n \geq a\tau$ satisfying*

$$\lambda \geq K\Big(\frac{a\tau}{n}\Big)^{\frac{1}{2-\vartheta+\vartheta p}} \tag{14}$$

we have

$$P^n\Big(D \in (X \times Y)^n : \mathcal{R}_{L,P}(f^{(1)}_{D,\lambda}) - \mathcal{R}^*_{L,P} < 2A_1(\lambda) + \lambda\Big) \;\geq\; 1 - e^{-\tau}\,.$$

To illustrate this second oracle inequality, let us again assume that L is the hinge loss, and that P satisfies Tsybakov's noise assumption (10). Then assumption (14) becomes

$$\lambda \geq K\Big(\frac{a\tau}{n}\Big)^{\frac{q+1}{q+pq+2}}\,, \tag{15}$$

which for $q = \infty$ reduces to $\lambda \geq K(\frac{a\tau}{n})^{\frac{1}{1+p}}$. Modulo constants, this is exactly the result from [2], but without the need of η being bounded away from 0 and 1. Moreover, unlike [2], our result holds for *all* $q \in [0, \infty]$, and hence it also provides a solution of another open problem of [2].

Let us finally compare the learning rates resulting from Theorem 2.3 and 2.4. To this end, we again restrict our considerations to the hinge loss L. In addition, we assume that there exists constants $c > 0$ and $\beta \in (0, 1]$ such that $A_2(\lambda) \leq c\lambda^{\beta}$ for all $\lambda > 0$. A simple calculation then shows that choosing $\lambda_n := n^{-\frac{2(q+1)}{(q+pq+2)(\beta+1)}}$ in (12) asymptotically minimizes (12) and the resulting learning rate is

$$n^{-\frac{2\beta(q+1)}{(q+pq+2)(\beta+1)}}\,. \tag{16}$$

On the other hand, for the lighter regularization, (15) shows that λ_n should asymptotically behave like $n^{-\frac{q+1}{q+pq+2}}$, which, by the relationship between A_2 and A_1 mentioned above, again yields the learning rate (16). In other words, the exponent of the regularization term does not have an effect on our learning rates, which seems reasonable if one recalls the fact that the regularization path is also independent of the exponent, see [13, Exercise 5.9]. We expect, that the same phenomenon holds, if one considers general exponents in the regularization term. Corresponding calculations should be straightforward but are clearly beyond the scope of this paper.

It is also worth mentioning that for the classical $\|\cdot\|^2_H$-regularization, the optimal choice of λ requires knowing p, q, *and* β, while for the lighter regularization, only

p and q need to be known. Of course, from a practical point of view, this does not make a big difference since typically q, and often also p, are not known, so that λ needs to be determined by, e.g., cross-validation approaches. From a theoretical point, however, it is interesting that for the lighter regularization the asymptotically optimal λ_n is independent of the approximation error (function) not only for $q = \infty$, as observed in [2], but also for $q < \infty$.

3 Proofs of the Oracle Inequalities

In order to prove the oracle inequalities we need to recall some results from [14]. To this end, we assume in the following that $q \in \{1, 2\}$ is fixed. We further define the function $C_\lambda : X \times Y \times H \to [0, \infty)$ by

$$C_\lambda(x, y, f) := \lambda \|f\|_H^q + L(y, f(x)) \,, \qquad x \in X,\, y \in Y,\, f \in H,$$

where $\lambda > 0$ is a regularization parameter. Note that this yields

$$\mathbb{E}_{(x,y)\sim P} C_\lambda(x, y, f) = \lambda \|f\|_H^q + \mathcal{R}_{L,P}(f) \,,$$

and following the arguments of [13, Chapter 5.1] it is not hard to see that the latter regularized risk not only has a unique minimizer if $q = 2$, but also in the case of $q = 1$. To avoid notational overload we denote this minimizer in both cases by $f_{P,\lambda}$, that is, in the case $q = 1$ we now write $f_{P,\lambda}$ and $f_{D,\lambda}$ rather than $f_{P,\lambda}^{(1)}$ and $f_{D,\lambda}^{(1)}$. Moreover, we need the induced classes

$$\mathcal{G}(\lambda) := \left\{ C_\lambda \circ f - C_\lambda \circ f_{P,\lambda} : f \in \lambda^{-1/q} B_H \right\} , \qquad \lambda > 0 \,,$$

where $C_\lambda \circ f := C_\lambda(\,\cdot\,, \cdot\,, f)$. Note that $\mathcal{R}_{L,P}(0) \le 1$ implies $f_{P,\lambda} \in \lambda^{-1/q} B_H$ for all distributions P on $X \times Y$, and hence the latter in particular holds for the empirical solutions $f_{D,\lambda}$. In other words, we have $C_\lambda \circ f_{D,\lambda} - C_\lambda \circ f_{P,\lambda} \in \mathcal{G}(\lambda)$ for all $D \in (X \times Y)^n$.

Furthermore recall that the modulus of continuity of the class $\mathcal{G}(\lambda)$ was defined by

$$\omega_{P,n}(\mathcal{G}(\lambda), \varepsilon) \;:=\; \mathbb{E}_{D\sim P^n} \Big(\sup_{\substack{f \in \mathcal{G}(\lambda), \\ \mathbb{E}_P f \le \varepsilon}} |\mathbb{E}_P f - \mathbb{E}_D f| \Big) \,,$$

where P is a probability measure on $X \times Y$. With the help of this modulus, [14, Theorem 3.1] establishes the following general oracle inequality

Theorem 3.1 *Adopt the above notations for fixed $q \in \{1, 2\}$ and $\lambda > 0$. Furthermore, assume that there exist constants $b, B \ge 0$, $\beta \in [0, 1]$, $w, W \ge 0$, and $\vartheta \in [0, 1]$ such that*

$$\|g\|_\infty \;\le\; b \left(\mathbb{E}_P g\right)^\beta + B \tag{17}$$

and

$$\mathbb{E}_P g^2 \;\le\; \Big(b \big(\mathbb{E}_P g\big)^\beta + B \Big)^{2-\vartheta} \Big(w \big(\mathbb{E}_P g\big)^\vartheta + W \Big) \tag{18}$$

for all $g \in \mathcal{G}(\lambda)$. Then for all $n \geq 1$, $\tau \geq 1$ and $\varepsilon > 0$ satisfying

$$\varepsilon \;\geq\; 3\omega_{P,n}(\mathcal{G}(\lambda),\varepsilon) + \sqrt{\frac{2\tau(b\varepsilon^{\beta}+B)^{2-\vartheta}(w\varepsilon^{\vartheta}+W)}{n}} + \frac{2\tau\big(b\varepsilon^{\beta}+B\big)}{n} \tag{19}$$

we have

$$P^n\Big(D \in (X\times Y)^n : \lambda\|f_{D,\lambda}\|_H^q + \mathcal{R}_{L,P}(f_{D,\lambda}) < A_q(\lambda) + \varepsilon\Big) \;\geq\; 1 - e^{-\tau}\,.$$

Let us now use the above general theorem to prove the two oracle inequalities presented in the previous section.

The case $\mathbf{q = 2}$

In the standard SVM case $q = 2$, the bounds (17) and (18) were guaranteed by [14, Lemma 4.1] and [14, Lemma 4.2], respectively. For the sake of convenience we recall both results:

Lemma 3.2 *For $0 < \lambda \leq 1$, and $f \in \lambda^{-\frac{1}{2}}B_H$ we define $g_f := C_\lambda \circ f - C_\lambda \circ f_{P,\lambda}$. Then we have $g_f \in \mathcal{G}(\lambda)$ and the following two bounds hold:*

$$\begin{aligned}
\|g_f\|_\infty &\leq 3\left(\frac{\mathbb{E}_P g_f}{\lambda}\right)^{1/2} + \left(\frac{A_2(\lambda)}{\lambda}\right)^{1/2} + 2\,,\\
\|f\|_H &\leq \left(\frac{A_2(\lambda) + \mathbb{E}_P g_f}{\lambda}\right)^{1/2}.
\end{aligned}$$

Lemma 3.3 *Let P be a distribution on $X\times Y$ and suppose that there exist constants $c \geq 1$ and $\vartheta \in [0,1]$ such that the variance bound assumption (9) is satisfied for some $0 < \lambda < 1$ and all $f \in \lambda^{-\frac{1}{2}}B_H$. Then for all $g \in \mathcal{G}(\lambda)$ we have*

$$\mathbb{E}_P g^2 \leq 16c\left(\left(\frac{\mathbb{E}_P g}{\lambda}\right)^{1/2} + \left(\frac{A_2(\lambda)}{\lambda}\right)^{1/2} + 1\right)^{2-\vartheta}\left((\mathbb{E}_P g)^{\vartheta} + 2A_2^{\vartheta}(\lambda)\right).$$

From these two lemmas it is easy to conclude that we may set $\beta := 1/2$, $b := 3\lambda^{-1/2}$, $B := (\frac{A_2(\lambda)}{\lambda})^{1/2}+2$, $w := 16c$, and $W := 32cA_2^{\vartheta}(\lambda)$ in Theorem 3.1. To apply the latter, it thus remains to find an upper bound on the modulus $\omega_{P,n}(\mathcal{G}(\lambda),\varepsilon)$.

Our next goal is to establish such an upper bound if we have a bound on certain random entropy numbers. Let us begin by recalling Rademacher averages. To this end, we fix a probability space $(\Theta,\mathcal{C},\nu)$, and a Rademacher sequence $\varepsilon_1,\ldots,\varepsilon_n$, that is, a sequence of i.i.d. random variables $\varepsilon_i : \Theta \to \{-1,1\}$ satisfying $\nu(\varepsilon_i = 1) = \nu(\varepsilon_i = -1) = 1/2$ for all $i = 1,\ldots,n$. Now let Z be a non-empty set equipped with some σ-algebra and $\mathcal{L}_0(Z)$ be the corresponding set of all measurable functions $g : Z \to \mathbb{R}$. Given a non-empty $\mathcal{G} \subset \mathcal{L}_0(Z)$, a Rademacher sequence $\varepsilon_1,\ldots,\varepsilon_n$, and a finite sequence $D := (z_1,\ldots,z_n) \in Z^n$, the n-th empirical Rademacher average of $\mathcal{G}$ is defined by

$$\mathrm{Rad}_D(\mathcal{G},n) := \mathbb{E}_\nu \sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(z_i)\right|.$$

It is well-known that *symmetrization*, see e.g. [17], makes it possible to bound the modulus of continuity by expected Rademacher averages. Namely we have

$$\omega_{P,n}(\mathcal{G}(\lambda),\varepsilon) \le 2\,\mathbb{E}_{D\sim P^n}\mathrm{Rad}_D(\mathcal{G}_\varepsilon,n)\,, \tag{20}$$

where

$$\mathcal{G}_\varepsilon := \{g\in\mathcal{G}(\lambda) : \mathbb{E}_P g \le \varepsilon\}\,.$$

In view of Theorem 3.1 it thus suffices to find a bound on the expected Rademacher averages of $\mathcal{G}_\varepsilon$. The classical way to obtain such a bound uses Dudley's chaining argument, see [5], [7], and Chapter 2.2 in [17], together with the covering numbers of $\mathcal{G}_\varepsilon$ with respect to $L_2(\mathrm{D})$. For our purposes, however, it is more convenient to use entropy numbers instead of covering numbers. Fortunately, Dudley's chaining argument works with entropy numbers as well as with covering numbers, see [13, Theorems 7.13 and 7.16]. In order to recall the latter result, we define the (dyadic) entropy numbers of a subset $A\subset H$ of a Hilbert space H by

$$e_i(A,H) \;:=\; \inf\biggl\{\varepsilon>0 : \exists\, x_1,\ldots,x_{2^{i-1}}\in A \text{ such that } A\subset \bigcup_{j=1}^{2^{i-1}}\bigl(x_j+\varepsilon B_H\bigr)\biggr\}\,.$$

Now the following version of Dudley's chaining whose proof can be found in Chapter 7.3 of [13] bounds empirical Rademacher averages by entropy numbers.

Theorem 3.4 *For every non-empty set $\mathcal{G}\subset\mathcal{L}_0(Z)$ and every finite sequence $D:=(z_1,\ldots,z_n)\in Z^n$, we have*

$$\mathrm{Rad}_D(\mathcal{G},n) \le \sqrt{\frac{\ln 16}{n}}\biggl(\sum_{i=1}^\infty 2^{i/2}\, e_{2^i}\bigl(\mathcal{G}\cup\{0\},L_2(\mathrm{D})\bigr) + \sup_{g\in\mathcal{G}}\|g\|_{L_2(\mathrm{D})}\biggr)\,.$$

Using Theorem 3.4 and an imposed bound on the average entropy numbers, the following theorem provides a bound on expected Rademacher averages. Its proof follows the ideas of [6] and can again be found in Chapter 7.3 of [13].

Theorem 3.5 *Let $\mathcal{G}\subset\mathcal{L}_0(Z)$ be a non-empty set and P be a distribution on Z. Suppose that there exist constants $B\ge 0$ and $\sigma\ge 0$ such that $\|h\|_\infty\le B$ and $\mathbb{E}_P h^2\le\sigma$ for all $h\in\mathcal{G}$. Furthermore, assume that for a fixed $n\ge 1$ there exist constants $p\in(0,1)$ and $a\ge B^{2p}$ such that*

$$\mathbb{E}_{D\sim P^n}\, e_i(\mathcal{G},L_2(\mathrm{D})) \;\le\; a^{\frac{1}{2p}}\, i^{-\frac{1}{2p}}\,, \qquad\qquad i\ge 1. \tag{21}$$

Then there exist constants $C_1(p)>0$ and $C_2(p)>0$ depending only on p such that

$$\mathbb{E}_{D\sim P^n}\mathrm{Rad}_D(\mathcal{G},n) \le \max\Bigl\{C_1(p)\,a^{\frac12}\sigma^{\frac{1-p}{2}}n^{-\frac12},\, C_2(p)\,a^{\frac{1}{1+p}}B^{\frac{1-p}{1+p}}n^{-\frac{1}{1+p}}\Bigr\}\,.$$

To apply this general result to the sets $\mathcal{G}_\varepsilon$ we finally need the quantity

$$\Lambda_2(\varepsilon,\lambda) := \varepsilon + A_2(\lambda) + \lambda\,, \qquad\qquad \varepsilon>0,\ \lambda>0. \tag{22}$$

Now the upper bound on the expected Rademacher averages reads as follows:

Lemma 3.6 *Let $n \in \mathbb{N}$, and assume that there are constants $a \geq 1$ and $p \in (0,1)$ such that*

$$\mathbb{E}_{D_X \sim P_X^n} e_i(\mathrm{id} : H \to L_2(\mathrm{D}_X)) \;\leq\; a^{\frac{1}{2p}} i^{-\frac{1}{2p}}\,, \qquad\qquad i \geq 1. \tag{23}$$

Then there exists a constant $c_p > 0$ depending only on p such that for all distributions P on $X \times Y$, all $\lambda \in (0,1]$, $\varepsilon \in (0,1]$, and all $\tau_\varepsilon \geq \sup_{g \in \mathcal{G}_\varepsilon} \mathbb{E}_P g^2$ we have

$$\mathbb{E}_{D \sim P^n} \mathrm{Rad}_D(\mathcal{G}_\varepsilon, n) \leq c_p \max\left\{ \tau_\varepsilon^{\frac{1-p}{2}} \left(\frac{\Lambda_2(\varepsilon,\lambda)}{\lambda}\right)^{\frac{p}{2}} \left(\frac{a}{n}\right)^{\frac{1}{2}}, \left(\frac{\Lambda_2(\varepsilon,\lambda)}{\lambda}\right)^{\frac{1}{2}} \left(\frac{a}{n}\right)^{\frac{1}{1+p}} \right\}.$$

Proof: Lemma 3.2 shows that for all $f \in \lambda^{-1/2} B_H$ with $g_f := C_\lambda \circ f - C_\lambda \circ f_{P,\lambda} \in \mathcal{G}_\varepsilon$ we have

$$\|f\|_H \leq \left(\frac{A_2(\lambda) + \varepsilon}{\lambda}\right)^{1/2} =: \Lambda\,.$$

Let us therefore write $\tilde{\mathcal{G}}_\varepsilon := \{\lambda \|f\|_H^2 + L \circ f : f \in \Lambda B_H\}$ and $\mathcal{H} := \{L \circ f : f \in \Lambda B_H\}$. Now observe that $\lambda \|f\|_H^2 \leq 2$ for all $f \in \Lambda B_H$, and hence the additivity of the entropy numbers, see [3, page 21], and their quasi-injectivity, see [3, (1.1.3) & (1.1.4)], together with the Lipschitz continuity of L yields

$$\begin{aligned} e_{2i-1}\bigl(\mathcal{G}_\varepsilon, L_2(\mathrm{D})\bigr) \leq 2 e_{2i-1}\bigl(\tilde{\mathcal{G}}_\varepsilon, L_2(\mathrm{D})\bigr) &\leq 2 e_i\bigl([0,2], |\cdot|\bigr) + 2 e_i\bigl(\mathcal{H}, L_2(\mathrm{D}_X)\bigr) \\ &\leq 2^{2-i} + 4 e_i\bigl(\Lambda B_H, L_2(\mathrm{D}_X)\bigr) \end{aligned}$$

for all $i \geq 1$ and all $D \in (X \times Y)^n$. Averraging over D and using (23), we thus obtain

$$\mathbb{E}_{D \sim P^n} e_{2i-1}\bigl(\mathcal{G}_\varepsilon, L_2(\mathrm{D})\bigr) \leq 2^{2-i} + 4\Lambda a^{\frac{1}{2p}} i^{-\frac{1}{2p}} \leq \tilde{c}_p (\Lambda^2 + 1)^{\frac{1}{2}} a^{\frac{1}{2p}} i^{-\frac{1}{2p}}$$

for a suitable constant $\tilde{c}_p$ only depending on p. From this it is straightforward to conclude that

$$\mathbb{E}_{D \sim P^n} e_i\bigl(\mathcal{G}_\varepsilon, L_2(\mathrm{D})\bigr) \leq c_p (\Lambda^2 + 1)^{\frac{1}{2}} a^{\frac{1}{2p}} i^{-\frac{1}{2p}}$$

for all $i \geq 1$, where c_p is another constant only depending on p. Now observe that, for $f \in \lambda^{-1/2} B_H$ with $g_f \in \mathcal{G}_\varepsilon$, we have $\|L \circ f\|_\infty \leq 1 + \|f\|_\infty \leq 1 + \|f\|_H \leq 1 + \Lambda$ and $\lambda \|f\|_H^2 \leq 1$. From this it is easy to conclude that $\|g_f\|_\infty \leq \Lambda + 3 =: B$ for all $g_f \in \mathcal{G}_\varepsilon$. Assuming without loss of generality that $c_p \geq \sqrt{18}$ we hence find for $\tilde{a} := c_p^{2p} (\Lambda^2 + 1)^p a$ that $\tilde{a} \geq B^{2p}$. Applying Theorem 3.5 and $\Lambda^2 + 1 = \lambda^{-1} \Lambda_2(\varepsilon, \lambda)$ then yields the assertion. ∎

Proof of Theorem 2.3: As already indicated after Lemma 3.3 we will apply Theorem 3.1 with $\beta := 1/2$, $b := 3\lambda^{-1/2}$, $B := (\frac{A_2(\lambda)}{\lambda})^{1/2} + 2$, $w := 16c$, and $W := 32 c A_2^\vartheta(\lambda)$. To do so, we first observe that with these definitions we have

$$b\varepsilon^\beta + B \leq 3\lambda^{-1/2}\bigl(\varepsilon^{1/2} + A_2^{1/2}(\lambda) + \lambda^{1/2}\bigr) \leq 3\sqrt{\frac{3\Lambda_2(\varepsilon,\lambda)}{\lambda}}$$

and

$$w\varepsilon^\vartheta + W \leq 32c\bigl(\varepsilon^\vartheta + A_2^\vartheta(\lambda)\bigr) \leq 64\, c\, \Lambda_2^\vartheta(\varepsilon,\lambda)\,,$$

where $\Lambda_2(\varepsilon,\lambda)$ is defined by (22). Moreover, Lemma 3.3 shows that all $g \in \mathcal{G}_\varepsilon$ satisfy

$$\begin{aligned}
\mathbb{E}_P g^2 &\leq 16c\biggl(\Bigl(\frac{\varepsilon}{\lambda}\Bigr)^{1/2} + \Bigl(\frac{A_2(\lambda)}{\lambda}\Bigr)^{1/2} + 1\biggr)^{2-\vartheta}\bigl(\varepsilon^\vartheta + 2A_2^\vartheta(\lambda)\bigr) \\
&\leq 16\sqrt{3}c\biggl(\frac{\varepsilon + A_2(\lambda) + \lambda}{\lambda}\biggr)^{1-\vartheta/2} 4\bigl(\varepsilon + A_2(\lambda)\bigr)^\vartheta \\
&\leq 64\sqrt{3}c\lambda^{\vartheta/2-1}\Lambda_2^{1+\vartheta/2}(\varepsilon,\lambda)\,.
\end{aligned}$$

For $\tau_\varepsilon := 64\sqrt{3}c\lambda^{\vartheta/2-1}\Lambda_2^{1+\vartheta/2}(\varepsilon,\lambda)$, Lemma 3.6 together with (20) then yields a constant C_p only depending on p and c such that

$$\omega_{P,n}(\mathcal{G}(\lambda),\varepsilon) \leq C_p\,\max\biggl\{\lambda^{\frac{\vartheta-\vartheta p-2}{4}}\Lambda_2^{\frac{2+\vartheta-\vartheta p}{4}}(\varepsilon,\lambda)\Bigl(\frac{a}{n}\Bigr)^{\frac12}, \biggl(\frac{\Lambda_2(\varepsilon,\lambda)}{\lambda}\biggr)^{\frac12}\Bigl(\frac{a}{n}\Bigr)^{\frac{1}{1+p}}\biggr\}\,.$$

Let us now restrict our considerations to ε that satisfy $\varepsilon \geq A_2(\lambda) + \lambda$. Then we obviously have $\Lambda_2(\varepsilon,\lambda) \leq 2\varepsilon$ and hence Condition (19) is satisfied for such ε, if

$$\varepsilon \;\geq\; \tilde{C}_p\max\biggl\{\lambda^{\frac{\vartheta-\vartheta p-2}{4}}\varepsilon^{\frac{2+\vartheta-\vartheta p}{4}}\Bigl(\frac{a}{n}\Bigr)^{\frac12}, \Bigl(\frac{\varepsilon}{\lambda}\Bigr)^{\frac12}\Bigl(\frac{a}{n}\Bigr)^{\frac{1}{1+p}}, \lambda^{\frac{\vartheta-2}{4}}\varepsilon^{\frac{2+\vartheta}{4}}\Bigl(\frac{\tau}{n}\Bigr)^{\frac12}, \Bigl(\frac{\varepsilon}{\lambda}\Bigr)^{\frac12}\frac{\tau}{n}\biggr\}\,,$$

where $\tilde{C}_p$ is another constant only depending on p and c. Simple algebraic transformations then reveal that the latter is satisfied if

$$\varepsilon \;\geq\; K\lambda^{-1}\max\biggl\{\Bigl(\frac{a}{n}\Bigr)^{\frac{2}{2-\vartheta+\vartheta p}}, \Bigl(\frac{a}{n}\Bigr)^{\frac{2}{2+p}}, \Bigl(\frac{\tau}{n}\Bigr)^{\frac{2}{2-\vartheta}}, \Bigl(\frac{\tau}{n}\Bigr)^2\biggr\}\,,$$

where K is yet another constant only depending on p and c. Applying Theorem 3.1 and $n \geq \tau$ then yields the assertion. ∎

The case $\mathbf{q = 1}$

In view of the proof of Theorem 2.3 we first need to find analogues for Lemmas 3.2, 3.3, and 3.6. Let us begin with an analogue for the first:

Lemma 3.7 *For $0 < \lambda \leq 1$, and $f \in \lambda^{-1}B_H$ we define $g := C_\lambda \circ f - C_\lambda \circ f_{P,\lambda}$, where q is assumed to equal 1. Then we have $g \in \mathcal{G}(\lambda)$ and the following two bounds hold:*

$$\begin{aligned}
\|g\|_\infty &\leq 4\cdot\frac{\mathbb{E}_P g + A_1(\lambda) + \lambda}{\lambda}\,, \\
\|f\|_H &\leq \frac{\mathbb{E}_P g + A_1(\lambda)}{\lambda}.
\end{aligned}$$

Proof: Let us fix an $f \in H$. Then we have

$$\begin{aligned}
\lambda\|f\|_H &\leq \lambda\|f\|_H + \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} \\
&= \lambda\|f_{P,\lambda}\|_H + \mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}^*_{L,P} + \mathbb{E}_P g \\
&= A_1(\lambda) + \mathbb{E}_P g\,,
\end{aligned}$$

and hence the second assertion follows. In order to show the first assertion, we first observe that the Lipschitz continuity of L together with $L(y,0) \leq 1$ implies $L(y,t) \leq 1 + |t|$ for all $y \in Y$ and $t \in \mathbb{R}$. By $\|\cdot\|_\infty \leq \|\cdot\|_H$ and the already proved second assertion, we consequently obtain

$$\begin{aligned} \big\|\lambda\|f\|_H + L \circ f\big\|_\infty \leq \lambda\|f\|_H + 1 + \|f\|_\infty &\leq A_1(\lambda) + \mathbb{E}_P g + 1 + \frac{A_1(\lambda) + \mathbb{E}_P g}{\lambda} \\ &\leq 2 \cdot \frac{A_1(\lambda) + \lambda + \mathbb{E}_P g}{\lambda}\,, \end{aligned}$$

where in the last step we used $0 < \lambda \leq 1$. Since this inequality holds for all $f \in H$, we then obtain the assertion. ∎

The next lemma establishes a variance bound similar to Lemma 3.3.

Lemma 3.8 *Let P be a distribution on $X \times Y$ and suppose that there exist constants $c \geq 1$ and $\vartheta \in [0,1]$ such that the variance bound assumption (9) is satisfied for some $0 < \lambda < 1$ and all $f \in \lambda^{-1} B_H$. Then for all $g \in \mathcal{G}(\lambda)$ we have*

$$\mathbb{E}_P g^2 \leq 4\,c \left(4 \cdot \frac{\mathbb{E}_P g + A_1(\lambda) + \lambda}{\lambda}\right)^{2-\vartheta} \big((\mathbb{E}_P g)^\vartheta + 2A_1^\vartheta(\lambda)\big)\,.$$

Proof: We use the shorthand notation $\mathbb{E}$ for $\mathbb{E}_P$ and $\|\cdot\|$ for $\|\cdot\|_H$. For $g \in \mathcal{G}(\lambda)$, we begin by picking an $f \in \lambda^{-1} B_H$ with $g = C_\lambda \circ f - C_\lambda \circ f_{P,\lambda}$. Now observe that

$$\begin{aligned} \mathbb{E}g^2 &= \mathbb{E}\big(\lambda\|f\| - \lambda\|f_{P,\lambda}\| + L \circ f - L \circ f_{P,\lambda}\big)^2 \\ &\leq 2\mathbb{E}\big(\lambda\|f\| - \lambda\|f_{P,\lambda}\|\big)^2 + 2\mathbb{E}\big(L \circ f - L \circ f_{P,\lambda}\big)^2 \\ &\leq 2\lambda^2\|f\|^2 + 2\lambda^2\|f_{P,\lambda}\|^2 + 2\mathbb{E}\big(L \circ f - L \circ f_{P,\lambda}\big)^2 \\ &\leq 4\mathbb{E}\big(L \circ f - L \circ f^*_{L,P}\big)^2 + 4\mathbb{E}\big(L \circ f^*_{L,P} - L \circ f_{P,\lambda}\big)^2 + 2\lambda^2\|f\|^2 + 2\lambda^2\|f_{P,\lambda}\|^2. \end{aligned}$$

We write $C := \max\big(\|f\|_\infty + 1, \|f_{P,\lambda}\|_\infty + 1\big)$. Then the assumption (9) and $a^\vartheta + b^\vartheta \leq 2(a+b)^\vartheta$ for all $a, b \geq 0$, imply that

$$\begin{aligned} &\mathbb{E}\big(L \circ f - L \circ f^*_{L,P}\big)^2 + \mathbb{E}\big(L \circ f^*_{L,P} - L \circ f_{P,\lambda}\big)^2 \\ \leq\ & 2cC^{2-\vartheta}\Big(\mathbb{E}\big(L \circ f - L \circ f^*_{L,P}\big) + \mathbb{E}\big(L \circ f_{P,\lambda} - L \circ f^*_{L,P}\big)\Big)^\vartheta\,. \end{aligned}$$

Since $\lambda\|f\| \leq 1$, $\lambda\|f_{P,\lambda}\| \leq 1$, and $\vartheta \leq 1$ we hence obtain

$$\begin{aligned} \mathbb{E}g^2 &\leq 8cC^{2-\vartheta}\Big(\mathbb{E}\big(L\circ f - L\circ f^*_{L,P}\big) + \mathbb{E}\big(L\circ f_{P,\lambda} - L\circ f^*_{L,P}\big)\Big)^\vartheta + 2\lambda^2\|f\|^2 + 2\lambda^2\|f_{P,\lambda}\|^2 \\ &\leq 8cC^{2-\vartheta}\Big(\mathbb{E}\big(L\circ f - L\circ f^*_{L,P}\big) + \mathbb{E}\big(L\circ f_{P,\lambda} - L\circ f^*_{L,P}\big)\Big)^\vartheta + 4\Big(\lambda\|f\| + \lambda\|f_{P,\lambda}\|\Big)^\vartheta \\ &\leq 16cC^{2-\vartheta}\Big(\mathbb{E}\big(L\circ f - L\circ f^*_{L,P}\big) + \mathbb{E}\big(L\circ f_{P,\lambda} - L\circ f^*_{L,P}\big) + \lambda\|f\| + \lambda\|f_{P,\lambda}\|\Big)^\vartheta \\ &= 16cC^{2-\vartheta}\Big(\mathbb{E}g + 2\mathbb{E}\big(L\circ f_{P,\lambda} - L\circ f^*_{L,P}\big) + 2\lambda\|f_{P,\lambda}\|\Big)^\vartheta \\ &\leq 16cC^{2-\vartheta}\Big((\mathbb{E}g)^\vartheta + 2A_1^\vartheta(\lambda)\Big)\,. \end{aligned}$$

Consequently, it remains to bound C on the right hand side of this inequality. To that end, observe that Lemma 3.7 implies

$$\|f\|_\infty \leq \|f\|_H \leq \frac{\mathbb{E}g + A_1(\lambda)}{\lambda}$$

and

$$\|f_{P,\lambda}\|_\infty \leq \|f_{P,\lambda}\|_H \leq \frac{A_1(\lambda)}{\lambda} \leq \frac{\mathbb{E}g + A_1(\lambda)}{\lambda}\,,$$

and hence we can bound

$$C = \max\Big(\|f\|_\infty + 1, \|f_{P,\lambda}\|_\infty + 1\Big) \leq \frac{\mathbb{E}g + A_1(\lambda)}{\lambda} + 1\,.$$

Combining the estimates then yields the assertion. ∎

Let us finally establish a bound on the expected Rademacher averages of the sets $\mathcal{G}_\varepsilon$ for the case $q = 1$. To this end we write

$$\Lambda_1(\varepsilon, \lambda) := \varepsilon + A_1(\lambda) + \lambda\,. \tag{24}$$

Now the upper bound on the expected Rademacher averages reads as follows:

Lemma 3.9 *Let $n \in \mathbb{N}$, and assume that there are constants $a \geq 1$ and $p \in (0,1)$ such that (23) is satisfied. Then there exists a constant $c_p > 0$ depending only on p such that for all distributions P on $X \times Y$, all $\lambda \in (0,1]$, $\varepsilon \in (0,1]$, and all $\tau_\varepsilon \geq \sup_{g \in \mathcal{G}_\varepsilon} \mathbb{E}_P g^2$ we have*

$$\mathbb{E}_{D\sim P^n} \mathrm{Rad}_D(\mathcal{G}_\varepsilon, n) \leq c_p \max\left\{ \tau_\varepsilon^{\frac{1-p}{2}} \left(\frac{\Lambda_1(\varepsilon,\lambda)}{\lambda}\right)^p \left(\frac{a}{n}\right)^{\frac{1}{2}}, \frac{\Lambda_1(\varepsilon,\lambda)}{\lambda} \left(\frac{a}{n}\right)^{\frac{1}{1+p}} \right\}.$$

Proof: Lemma 3.7 shows that for all $f \in \lambda^{-1} B_H$ with $g_f := C_\lambda \circ f - C_\lambda \circ f_{P,\lambda} \in \mathcal{G}_\varepsilon$ we have

$$\|f\|_H \leq \frac{A_1(\lambda) + \varepsilon}{\lambda} =: \Lambda\,.$$

Let us therefore write $\tilde{\mathcal{G}}_\varepsilon := \{\lambda\|f\|_H + L \circ f : f \in \Lambda B_H\}$ and $\mathcal{H} := \{L \circ f : f \in \Lambda B_H\}$. Now observe that $\lambda\|f\|_H \leq 2$ for all $f \in \Lambda B_H$, and we find

$$\begin{aligned} e_{2i-1}\big(\mathcal{G}_\varepsilon, L_2(\mathrm{D})\big) \leq 2e_{2i-1}\big(\tilde{\mathcal{G}}_\varepsilon, L_2(\mathrm{D})\big) &\leq 2e_i\big([0,2], |\cdot|\big) + 2e_i\big(\mathcal{H}, L_2(\mathrm{D}_X)\big) \\ &\leq 2^{2-i} + 4e_i\big(\Lambda B_H, L_2(\mathrm{D}_X)\big) \end{aligned}$$

for all $i \geq 1$ and all $D \in (X \times Y)^n$ As in the proof of Lemma 3.6 we then conclude that

$$\mathbb{E}_{D\sim P^n} e_i\big(\mathcal{G}_\varepsilon, L_2(\mathrm{D})\big) \leq c_p(\Lambda + 1) a^{\frac{1}{2p}}\, i^{-\frac{1}{2p}}$$

for all $i \geq 1$, where c_p is a constant only depending on p. Now observe that, for $f \in \lambda^{-1} B_H$ with $g_f \in \mathcal{G}_\varepsilon$, we have $\|L \circ f\|_\infty \leq 1 + \|f\|_\infty \leq 1 + \|f\|_H \leq 1 + \Lambda$ and $\lambda\|f\|_H \leq 1$. From this it is easy to conclude that $\|g_f\|_\infty \leq \Lambda + 3 =: B$ for all $g_f \in \mathcal{G}_\varepsilon$. Assuming without loss of generality that $c_p \geq 3$, we hence find for $\tilde{a} := c_p^{2p}(\Lambda + 1)^{2p} a$ that $\tilde{a} \geq B^{2p}$. Applying Theorem 3.5 and $\Lambda + 1 = \lambda^{-1}\Lambda_1(\varepsilon, \lambda)$ then yields the assertion. ∎

Proof of Theorem 2.4: We will apply Theorem 3.1 with $\beta := 1$, $b := 4\lambda^{-1}$, $B := 4\frac{A_1(\lambda)+\lambda}{\lambda}$, $w := 4c$, and $W := 8cA_1^\vartheta(\lambda)$. To do so, we first observe that with these definitions we have

$$b\varepsilon^\beta + B = 4\lambda^{-1}\varepsilon + 4\cdot\frac{A_1(\lambda)+\lambda}{\lambda} = 4\cdot\frac{\Lambda_1(\varepsilon,\lambda)}{\lambda}$$

and

$$w\varepsilon^\vartheta + W = 4c\varepsilon^\vartheta + 8cA_1^\vartheta(\lambda) \le 16\,c\,\Lambda_1^\vartheta(\varepsilon,\lambda)\,,$$

where $\Lambda_1(\varepsilon,\lambda)$ is defined by (24). Moreover, Lemma 3.8 shows that all $g \in \mathcal{G}_\varepsilon$ satisfy

$$\begin{aligned}\mathbb{E}_P g^2 &\le 4\,c\left(4\cdot\frac{\varepsilon+A_1(\lambda)+\lambda}{\lambda}\right)^{2-\vartheta}\left(\varepsilon^\vartheta + 2A_1^\vartheta(\lambda)\right)\\ &\le 32c\left(\frac{\varepsilon+A_1(\lambda)+\lambda}{\lambda}\right)^{2-\vartheta}\left(\varepsilon+2A_1(\lambda)\right)^\vartheta\\ &\le 64c\lambda^{\vartheta-2}\Lambda_1^2(\varepsilon,\lambda)\,.\end{aligned}$$

For $\tau_\varepsilon := 64c\lambda^{\vartheta-2}\Lambda_1^2(\varepsilon,\lambda)$, Lemma 3.9 together with (20) then yields a constant C_p only depending on p and c such that

$$\omega_{P,n}(\mathcal{G}(\lambda),\varepsilon) \le C_p\,\max\left\{\lambda^{\frac{\vartheta-\vartheta p-2}{2}}\Lambda_1(\varepsilon,\lambda)\left(\frac{a}{n}\right)^{\frac12}, \frac{\Lambda_1(\varepsilon,\lambda)}{\lambda}\cdot\left(\frac{a}{n}\right)^{\frac{1}{1+p}}\right\}.$$

Let us now restrict our considerations to ε that satisfy $\varepsilon \ge A_1(\lambda)+\lambda$. Then we obviously have $\Lambda_1(\varepsilon,\lambda) \le 2\varepsilon$ and hence Condition (19) is satisfied for such ε, if

$$\varepsilon \;\ge\; \tilde{C}_p\max\left\{\lambda^{\frac{\vartheta-\vartheta p-2}{2}}\varepsilon\left(\frac{a}{n}\right)^{\frac12}, \frac{\varepsilon}{\lambda}\cdot\left(\frac{a}{n}\right)^{\frac{1}{1+p}}, \lambda^{\frac{\vartheta-2}{2}}\varepsilon\cdot\left(\frac{\tau}{n}\right)^{\frac12}, \frac{\varepsilon}{\lambda}\cdot\frac{\tau}{n}\right\},$$

where $\tilde{C}_p \ge 1$ is another constant only depending on p and c. Simple algebraic transformations then reveal that the latter is satisfied if (14) is satisfied for

$$K := \tilde{C}_p^{\frac{2}{2-\vartheta}}\,.$$

Applying Theorem 3.1 thus yields the assertion. ∎

4 Appendix

In this appendix, we present the proofs of Theorem 2.1 and Lemma 2.2. Note that Theorem 2.1 has been essentially established in [13, Chapter 7.5], while Lemma 2.2 is somewhat well-known for people familiar with s-numbers introduced below. Nonetheless we feel that these results are not accessible enough for the statistical learning theory community, so we decided to recompile their proofs in this appendix.

Let us begin by describing the connection between eigenvalues and entropy numbers for certain operators acting on a Hilbert space. To this end, let H_1 and H_2 be two (real) Hilbert spaces and $S : H_1 \to H_2$ be a bounded linear operator. We say that S is compact, if the closure of the image SB_{H_1} is a compact subset of H_2.

We further denote the adjoint of S by S^*, i.e., S^* is the operator which is uniquely determined by the relation

$$\langle Sx, y\rangle_{H_2} = \langle x, S^*y\rangle_{H_1}, \qquad x \in H_1,\, y \in H_2.$$

Recall that a bounded linear operator $T : H \to H$ is called self-adjoint if $T^* = T$, and it is called positive if $\langle Tx, x\rangle \geq 0$. Given a bounded linear operator $S : H_1 \to H_2$, it is elementary to see that S^*S and SS^* are self-adjoint and positive.

It is well-known that for compact, self-adjoint, and positive operators $T : H \to H$ there exist an at most countable orthonormal system $(e_i)_{i\in I}$ of H and a family $(\lambda_i(T))_{i\in I}$ such that $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ and

$$Tx = \sum_{i\in I} \lambda_i(T)\langle x, e_i\rangle e_i, \qquad x \in H. \tag{25}$$

Moreover, $\{\lambda_i(T) : i \in I\}$ is the set of non-zero eigenvalues of T. In the following, we assume that I is of the form $I = \{1, 2, \ldots, |I|\}$ if the cardinality $|I|$ of I is finite. In this case, we define $\lambda_i(T) := 0$ for all $i > |I|$. Moreover, if $|I| = \infty$, we assume without loss of generality that $I = \mathbb{N}$. In both cases, we call $(\lambda_i(T))_{i\geq 1}$ the *extended* sequence of eigenvalues of T.

Now observe that given a compact $S : H_1 \to H_2$, the operator $S^*S : H_1 \to H_1$ is compact, positive, and self-adjoint, and hence it enjoys a representation of the form (25) with non-negative eigenvalues. We write

$$s_i(S) := \sqrt{\lambda_i(S^*S)}\,, \qquad i \geq 1\,, \tag{26}$$

for the *singular numbers* of S, where $(\lambda_i(S^*S))_{i\geq 1}$ is the extended sequence of eigenvalues of S^*S. Recall that S^*S and SS^* have exactly the same non-zero eigenvalues with the same geometric multiplicities, and hence we find $s_i(S^*) = s_i(S)$ for all $i \geq 1$. Moreover, we have

$$s_i^2(S) = \lambda_i(S^*S) = s_i(S^*S)\,, \qquad i \geq 1, \tag{27}$$

where in the second equality we used the fact that for compact, positive, and self-adjoint $T : H \to H$ we have

$$s_i(T) = \sqrt{\lambda_i(T^*T)} = \sqrt{\lambda_i(T^2)} = \lambda_i(T)\,, \qquad i \geq 1. \tag{28}$$

Let us now consider another interesting property of the singular numbers. To this end, let $S : E \to F$ be a bounded linear operator acting between arbitrary Banach spaces E and F. For $i \geq 1$, its i-th *approximation number* is then defined by

$$a_i(S) := \inf\big\{\|S - A\| : A \in \mathcal{L}(E, F) \text{ with rank } A < i\big\}\,, \tag{29}$$

where $\mathcal{L}(E, F)$ denotes the set of all bounded linear operators between E and F. Obviously, $(a_i(S))_{i\geq 1}$ is decreasing, and if rank $S < \infty$, we also have $a_i(S) = 0$ for all $i >$ rank S. Moreover, by diagonalization (see, e.g., Section 2.11 of [9]), one can show that

$$s_i(S) = a_i(S) \tag{30}$$

for all compact $S \in \mathcal{L}(H_1, H_2)$ acting between Hilbert spaces and all $i \geq 1$. In other words, singular and approximation numbers coincide for compact operators on Hilbert spaces. Moreover, entropy numbers are also closely related to approximation numbers. Namely, Carl's inequality, see Theorem 3.1.2 in [3], states that for all $0 < p \leq \infty$ and $0 < q < \infty$ there exists a constant $c_{p,q} > 0$ such that

$$\sum_{i=1}^{m} i^{q/p-1} e_i^q(S) \leq c_{p,q} \sum_{i=1}^{m} i^{q/p-1} a_i^q(S) \tag{31}$$

for all bounded operators $S : E \to F$ acting between Banach spaces and all $m \geq 1$. In addition, [3, Theorem 3.1.2] shows that the same holds for the finite dimensional $\ell_{p,\infty}$ norms, that is, for all $0 < p < \infty$ there exists a constant only depending on p such that

$$\sup_{i \leq m} i^{1/p} e_i(S) \leq c_p \sup_{i \leq m} i^{1/p} a_i(S)\,. \tag{32}$$

In general, these inequalities cannot be inverted, but for Hilbert spaces H and compact operators $S : H_1 \to H_2$, we actually have the following strong inverse of the above inequalities:

$$a_i(S) \leq 2e_i(S)\,, \qquad\qquad i \geq 1\,. \tag{33}$$

For a proof we refer to p. 120 in [3]. With these preparation we can now prove Lemma 2.2:

Proof of Lemma 2.2: Let us define the operator $S_{k,\mu} : L_2(\mu) \to H$ by

$$S_{k,\mu} g(x) := \int_X k(x, x') g(x') d\mu(x')\,, \qquad\qquad g \in L_2(\mu),\, x \in X. \tag{34}$$

Then it is easy to show that $S_{k,\mu}$ is the adjoint of the inclusion $\mathrm{id} : H \to L_2(\mu)$, and hence we have $\mathrm{id} : H \to L_2(\mu) = S_{k,\mu}^*$. Consequently, we obtain $T_{k,\mu} = S_{k,\mu}^* \circ S_{k,\mu}$, and by combining (27), (30), and (33) we obtain

$$\lambda_i(T_{k,\mu}) = s_i(T_{k,\mu}) = s_i^2(S_{k,\mu}^*) = a_i^2(\mathrm{id} : H \to L_2(\mu)) \leq 4e_i^2(\mathrm{id} : H \to L_2(\mu))$$

for all $i \geq 1$. From this the inequality of the right hand side can be easily derived. Analogously, Carl's inequality (32) together with (30), (27), and (28) implies

$$c_p^{-1} \sup_{i \leq m} i^{1/p} e_i^2(\mathrm{id} : H \to L_2(\mu)) \leq \sup_{i \leq m} i^{1/p} a_i^2(S_{k,\mu}^*) = \sup_{i \leq m} i^{1/p} \lambda_i(T_{k,\mu})$$

for all $m \geq 1$. Letting $m \to \infty$ then yields the assertion. ∎

Lemma 2.2 shows that the entropy numbers $e_i(\mathrm{id} : H \to L_2(\mu))$ and the eigenvalues $\lambda_i(T_{k,\mu})$ are closely related to each other, and that this relation is *independent* of the measure μ. This suggests that Theorem 2.1 can be proved once we have established a relation between $\lambda_i(T_{k,\mu})$ and the average random eigenvalues $\mathbb{E}_{D \sim \mu^n} \lambda_i(T_{k,\mathrm{D}})$. Fortunately, a sufficient result in this direction has already been established by [11, 12] in the special case of continuous kernels over compact metric spaces. Moreover, [21] generalized this result to bounded measurable kernels with separable RKHSs. However, a close inspection of the proof of [21], see [13, Chapter 7.5], shows that the boundedness of the kernel k can be replaced by the weaker assumption $\|k\|_{L_2(\mu)} < \infty$. The corresponding result reads as follows:

Theorem 4.1 *Let k be a measurable kernel on X with separable RKHS H and μ be a probability measure on X such that $\|k\|_{L_2(\mu)} < \infty$. Then for all $m \geq 1$ we have*

$$\mathbb{E}_{D\sim\mu^n} \sum_{i=m}^{\infty} \lambda_i(T_{k,\mathrm{D}}) \leq \sum_{i=m}^{\infty} \lambda_i(T_{k,\mu})\,. \tag{35}$$

With the help of this theorem we can now establish a general inequality between $e_i(\mathrm{id} : H \to L_2(\mu))$ and $\mathbb{E}_{D\sim\mu^n} e_i(\mathrm{id} : H \to L_2(\mathrm{D}))$. As we will see below, the assertion of Theorem 2.1 is a simple consequence of this general inequality.

Theorem 4.2 *Let k be a measurable kernel on X with separable RKHS H and μ be a probability measure on X such that $\|k\|_{L_2(\mu)} < \infty$. Then for all $0 < p < \infty$ and all $0 < q \leq 2$ there exists a constant $c_{p,q} \geq 1$ only depending on p and q such that for all $n \geq 1$, $m \geq 1$, and $M := \min\{m, n\}$ we have*

$$\sum_{i=1}^{m} i^{q/p-1} \mathbb{E}_{D\sim\mu^n} e_i^q(\mathrm{id} : H \to L_2(\mathrm{D})) \leq c_{p,q} \sum_{i=1}^{M} i^{q/p-1} \biggl(\frac{1}{i} \sum_{j=i}^{\infty} e_j^2\bigl(\mathrm{id} : H \to L_2(\mu)\bigr) \biggr)^{q/2}.$$

Proof: Carl's inequality (31) shows that there exists a constant $c_{p,q} > 0$ such that for $m, n \geq 1$ and all $D \in X^n$ we have

$$\sum_{i=1}^{m} i^{q/p-1} e_i^q(S_{k,\mathrm{D}}^*) \leq c_{p,q} \sum_{i=1}^{m} i^{q/p-1} a_i^q(S_{k,\mathrm{D}}^*) = c_{p,q} \sum_{i=1}^{\min\{m,n\}} i^{q/p-1} a_i^q(S_{k,\mathrm{D}}^*)\,,$$

where in the last step we used that $n \geq \operatorname{rank} S_{k,D}^*$ implies $a_i(S_{k,\mathrm{D}}^*) = 0$ for all $i > n$. Moreover, for $M := \min\{m, n\}$ and $\tilde{M} := \lfloor (M+1)/2 \rfloor$, we have

$$\sum_{i=1}^{M} i^{q/p-1} a_i^q(S_{k,\mathrm{D}}^*) \leq \sum_{i=1}^{\tilde{M}} (2i-1)^{q/p-1} a_{2i-1}^q(S_{k,\mathrm{D}}^*) + \sum_{i=1}^{\tilde{M}} (2i)^{q/p-1} a_{2i}^q(S_{k,\mathrm{D}}^*)\,.$$

If $p \leq q$, the monotonicity of the approximation numbers thus yields

$$\sum_{i=1}^{M} i^{q/p-1} a_i^q(S_{k,\mathrm{D}}^*) \leq 2^{q/p} \sum_{i=1}^{M} i^{q/p-1} a_{2i-1}^q(S_{k,\mathrm{D}}^*)\,,$$

and if $p > q$ we analogously find

$$\sum_{i=1}^{M} i^{q/p-1} a_i^q(S_{k,\mathrm{D}}^*) \leq 2^{2+q/p} \sum_{i=1}^{M} i^{q/p-1} a_{2i-1}^q(S_{k,\mathrm{D}}^*)\,,$$

Using (30) and (26) we further see that $a_i^2(S_{k,\mathrm{D}}^*) = s_i^2(S_{k,\mathrm{D}}^*) = s_i(S_{k,\mathrm{D}}^* S_{k,\mathrm{D}}) = \lambda_i(T_{k,\mathrm{D}})$ for all $i \geq 1$ and $D \in X^n$. Since $q \leq 2$ we thus obtain

$$\begin{aligned}
\sum_{i=1}^{m} i^{q/p-1} \mathbb{E}_{D\sim\mu^n} e_i^q(S_{k,\mathrm{D}}^*) &\leq \tilde{c}_{p,q} \sum_{i=1}^{M} i^{q/p-1} \mathbb{E}_{D\sim\mu^n} a_{2i-1}^q(S_{k,\mathrm{D}}^*) \\
&\leq \tilde{c}_{p,q} \sum_{i=1}^{M} i^{q/p-1} \bigl(\mathbb{E}_{D\sim\mu^n} \lambda_{2i-1}(T_{k,\mathrm{D}})\bigr)^{q/2}\,,
\end{aligned}$$

where $\tilde{c}_{p,q} := 2^{2+q/p} c_{p,q}$. Now for each $D \in X^n$ the sequence $(\lambda_i(T_{k,\mathrm{D}}))_{i\geq 1}$ is monotonically decreasing and hence so is $(\mathbb{E}_{D\sim\mu^n}\lambda_i(T_{k,\mathrm{D}}))_{i\geq 1}$. By Theorem 4.1, we hence find

$$i\,\mathbb{E}_{D\sim\mu^n}\lambda_{2i-1}(T_{k,\mathrm{D}}) \leq \sum_{j=i}^{2i-1} \mathbb{E}_{D\sim\mu^n}\lambda_j(T_{k,\mathrm{D}}) \leq \sum_{j=i}^{\infty} \lambda_j(T_{k,\mu})$$

for all $i \geq 1$, and consequently we obtain

$$\sum_{i=1}^{M} i^{q/p-1}\big(\mathbb{E}_{D\sim\mu^n}\lambda_{2i-1}(T_{k,\mathrm{D}})\big)^{q/2} \leq \sum_{i=1}^{M} i^{q/p-1}\left(\frac{1}{i}\sum_{j=i}^{\infty}\lambda_j(T_{k,\mu})\right)^{q/2}.$$

Moreover, by (26), (27), and (30), we have

$$\lambda_j(T_{k,\mu}) = s_i(S^*_{k,\mu}\circ S_{k,\mu}) = s_i^2(S^*_{k,\mu}) = a_j^2(S^*_{k,\mu}) \leq 4e_j^2(S^*_{k,\mu})\,,$$

where in the last step we used (33). Combining the estimates above, we hence obtain

$$\sum_{i=1}^{m} i^{q/p-1}\mathbb{E}_{D\sim\mu^n} e_i^q(S^*_{k,\mathrm{D}}) \leq 2^q\tilde{c}_{p,q}\sum_{i=1}^{M} i^{q/p-1}\left(\frac{1}{i}\sum_{j=i}^{\infty} e_j^2(S^*_{k,\mu})\right)^{q/2},$$

i.e., we have also shown the assertion. ∎

Proof of Theorem 2.1: Since $0 < p < 2$, it is easy to see that there exists a constant $\tilde{c}_p$ such that

$$\frac{1}{i}\sum_{j=i}^{\infty} e_j^2(S^*_{k,\mu}) \leq a^2\cdot\frac{1}{i}\sum_{j=i}^{\infty} j^{-\frac{2}{p}} \leq \tilde{c}_p^2\, a^2\, i^{-\frac{2}{p}}$$

for all $i \geq 1$. Using $\frac{1}{p} - 1 > -1$, we hence find another constant $c'_p > 0$ such that for $m \geq 1$ we have

$$\sum_{i=1}^{m} i^{\frac{2}{p}-1}\left(\frac{1}{i}\sum_{j=i}^{\infty} e_j^2(S^*_{k,\mu})\right)^{1/2} \leq \tilde{c}_p\, a\sum_{i=1}^{m} i^{\frac{1}{p}-1} \leq c'_p\, a\, m^{\frac{1}{p}}\,. \tag{36}$$

Furthermore, for $\tilde{m} := \lfloor (m+1)/2 \rfloor$, the monotonicity of the entropy numbers yields

$$\tilde{m}^{\frac{2}{p}}\,\mathbb{E}_{D\sim\mu^n} e_m(S^*_{k,\mathrm{D}}) \leq \sum_{i=\tilde{m}}^{m} i^{\frac{2}{p}-1}\,\mathbb{E}_{D\sim\mu^n} e_i(S^*_{k,\mathrm{D}}) \leq \sum_{i=1}^{m} i^{\frac{2}{p}-1}\,\mathbb{E}_{D\sim\mu^n} e_i(S^*_{k,\mathrm{D}})\,,$$

and since $m/2 \leq \lfloor (m+1)/2 \rfloor = \tilde{m}$, we hence obtain

$$\mathbb{E}_{D\sim\mu^n} e_m(S^*_{k,\mathrm{D}}) \leq 4^{1/p} m^{-\frac{2}{p}} \sum_{i=1}^{m} i^{\frac{2}{p}-1}\,\mathbb{E}_{D\sim\mu^n} e_i(S^*_{k,\mathrm{D}})\,.$$

Combining this estimate with (36) and Theorem 4.2 for $\tilde{p} := p/2$ and $q := 1$ then yields first assertion. ∎

Although not needed for the analysis of this paper, we finally like to mention another corollary of Theorem 4.2.

Corollary 4.3 *Let k be a measurable kernel on X with separable RKHS H and μ be a probability measure on X such that $\|k\|_{L_2(\mu)} < \infty$. Then for all $0 < p < 2$ there exists a constant $c_p \geq 1$ only depending on p such that for all $n \geq 1$ we have*

$$\sum_{i=1}^{\infty} i^{2/p-1} \mathbb{E}_{D\sim\mu^n} e_i^2(\mathrm{id} : H \to L_2(\mathrm{D})) \leq c_p \sum_{i=1}^{\infty} i^{2/p-1} e_i^2(\mathrm{id} : H \to L_2(\mu)) \,.$$

Proof: For $q = 2$ the right-hand side of the inequality of Theorem 4.2 becomes

$$\sum_{i=1}^{M} i^{q/p-1} \biggl(\frac{1}{i} \sum_{j=i}^{\infty} e_j^2(S_{k,\mu}^*) \biggr)^{q/2} = \sum_{i=1}^{M} i^{2/p-2} \sum_{j=i}^{\infty} e_j^2(S_{k,\mu}^*) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} b_{i,j} \,,$$

where $b_{i,j} := 0$ if $i > \min\{j, M\}$ and $b_{i,j} := i^{2/p-2} e_j^2(S_{k,\mu}^*)$ otherwise. Moreover, rearranging the sums and using $p < 2$ yields a constant c_p such that

$$\begin{aligned}
\sum_{j=1}^{\infty} \sum_{i=1}^{\infty} b_{i,j} &= \sum_{j=1}^{M} \sum_{i=1}^{j} i^{2/p-2} e_j^2(S_{k,\mu}^*) + \sum_{j=M+1}^{\infty} \sum_{i=1}^{M} i^{2/p-2} e_j^2(S_{k,\mu}^*) \\
&\leq c_p \sum_{j=1}^{M} j^{2/p-1} e_j^2(S_{k,\mu}^*) + c_p \sum_{j=M+1}^{\infty} M^{2/p-1} e_j^2(S_{k,\mu}^*) \\
&\leq c_p \sum_{j=1}^{\infty} j^{2/p-1} e_j^2(S_{k,\mu}^*) \,.
\end{aligned}$$

Applying Theorem 4.2 then yields the assertion. ∎

References

[1] P. L. Bartlett, O. Bousquet, and S. Mendelson. Localized Rademacher complexities. In J. Kivinen and R. H. Sloan, editors, *Proceedings of the 15th Annual Conference on Learning Theory*, pages 44–58. Springer, New York, 2002.

[2] G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *Ann. Statist.*, 36:489–531, 2008.

[3] B. Carl and I. Stephani. *Entropy, Compactness and the Approximation of Operators*. Cambridge University Press, Cambridge, 1990.

[4] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, 2000.

[5] R. M. Dudley. The sizes of compact subsets of Hilbert spaces and continuity of Gaussian processes. *J. Funct. Anal.*, 1:290–330, 1967.

[6] S. Mendelson. Improving the sample complexity using global data. *IEEE Trans. Inform. Theory*, 48:1977–1991, 2002.

[7] S. Mendelson. A few notes on statistical learning theory. In S. Mendelson and A. Smola, editors, *Advanced Lectures on Machine Learning: Machine Learning Summer School 2002, Canberra, Australia*, pages 1–40. Springer, Berlin, 2003.

[8] S. Mendelson. On the performance of kernel classes. *J. Mach. Learn. Res.*, 4:759–771, 2003.

[9] A. Pietsch. *Eigenvalues and s-Numbers*. Geest & Portig K.-G., Leipzig, 1987.

[10] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[11] J. Shawe-Taylor, C. K I. Williams, N. Cristianini, and J. Kandola. On the eigenspectrum of the Gram matrix and its relationship to the operator eigenspectrum. In N. Cesa-Bianchi, M. Numao, and R. Reischuk, editors, *Algorithmic Learning Theory, 13th International Conference*, pages 23–40. Springer, New York, 2002.

[12] J. Shawe-Taylor, C. K. I. Williams, N. Cristianini, and J. Kandola. On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Trans. Inform. Theory*, 51:2510–2522, 2005.

[13] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.

[14] I. Steinwart, D. Hush, and C. Scovel. A new concentration result for regularized risk minimizers. In E. Giné, V. Koltchinskii, W. Li, and J. Zinn, editors, *High Dimensional Probability IV*, pages 260–275. Institute of Mathematical Statistics, Beachwood, OH, 2006.

[15] I. Steinwart and C. Scovel. Fast rates for support vector machines. In P. Auer and R. Meir, editors, *Proceedings of the 18th Annual Conference on Learning Theory*, pages 279–294. Springer, New York, 2005.

[16] I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.*, 35:575–607, 2007.

[17] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

[18] R. Williamson, A. J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Trans. Inform. Theory*, 47:2516–2532, 2001.

[19] Q. Wu, Y. Ying, and D.-X. Zhou. Multi-kernel regularized classifiers. *J. Complexity*, 23:108–134, 2007.

[20] D. X. Zhou. The covering number in learning theory. *J. Complexity*, 18:739–767, 2002.

[21] L. Zwald, O. Bousquet, and G. Blanchard. Statistical properties of kernel principal component analysis. In J. Shawe-Taylor and Y. Singer, editors, *Proceedings of the 17th Annual Conference on Learning Theory*, pages 594–608. Springer, New York, 2004.